CAUSAL INFERENCE

FOR SCIENCE AND DECISIONS

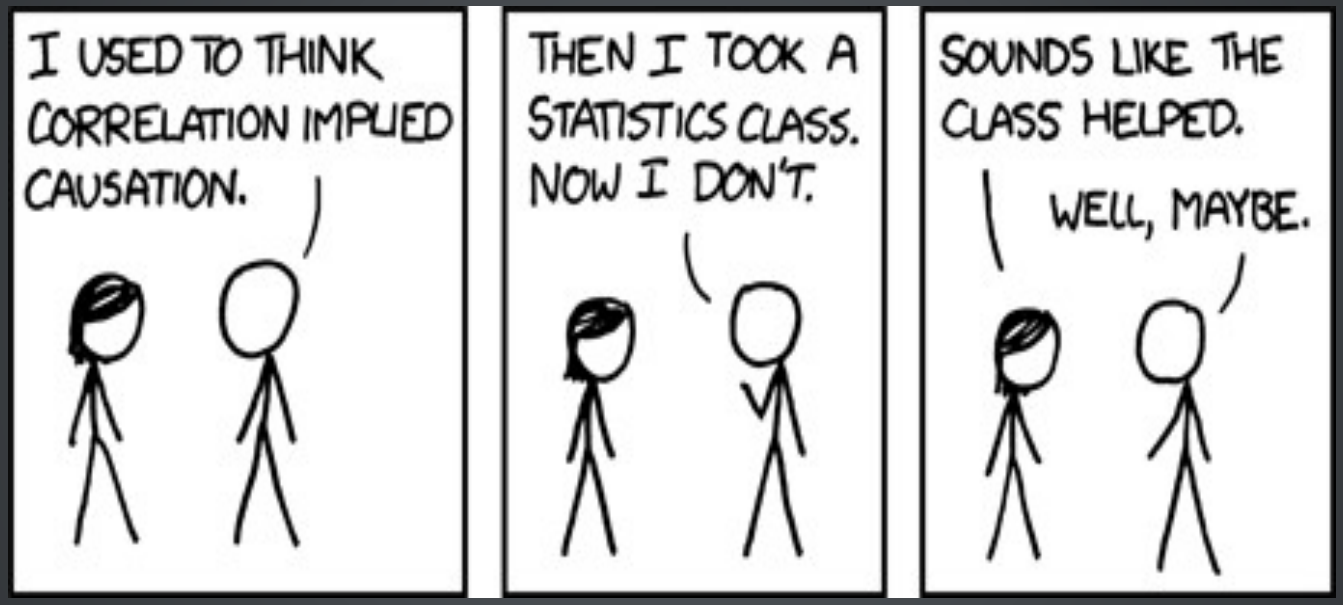
OUTLINE

Associations and Causal Graphs

Why do causal inference?

Experimental Design

OBLIGATORY STATISTICS JOKE



ASSOCIATIONS
(X_i, Y_i)

(smoking, cancer)
(running the ball, winning games)
(facebook fan, makes purchase)
(shown ad, makes purchase)
(toothbrushing, heart disease)
(SAT scores, success in college)

POSSIBLE RELATIONSHIPS

- $\Pr(Y_i \mid X_i) = \Pr(Y_i)$ (independence)
- $\Pr(Y_i \mid X_i) \neq \Pr(Y_i)$ (dependence)

Dependence is useful. It's how we build predictive models.

POSSIBLE CAUSAL RELATIONSHIPS

1. $X_i \leftarrow Y_i$
2. $X_i \rightarrow Y_i$
3. No causal relationship (but could still be correlated!)

Notice how a dependency doesn't say which direction the causal relationship is.

CORRELATION WITHOUT CAUSATION

INTRODUCING Z_i

- $X_i \rightarrow Z_i \rightarrow Y_i$ (chain)
- $X_i \leftarrow Z_i \leftarrow Y_i$ (chain)
- $X_i \leftarrow Z_i \rightarrow Y_i$ (fork)
- $X_i \rightarrow Z_i \leftarrow Y_i$ (collider)

CHAINS

X_i DOESN'T CAUSE Y_i, IT CAUSES THE CAUSE.

Example: McDonald's opening doesn't cause obesity, it causes overeating... which causes obesity

FORKS

Z_i CAUSES BOTH X_i AND Y_i.

Example: High natural ability causes good SAT scores and success in college.

COLLIDERS

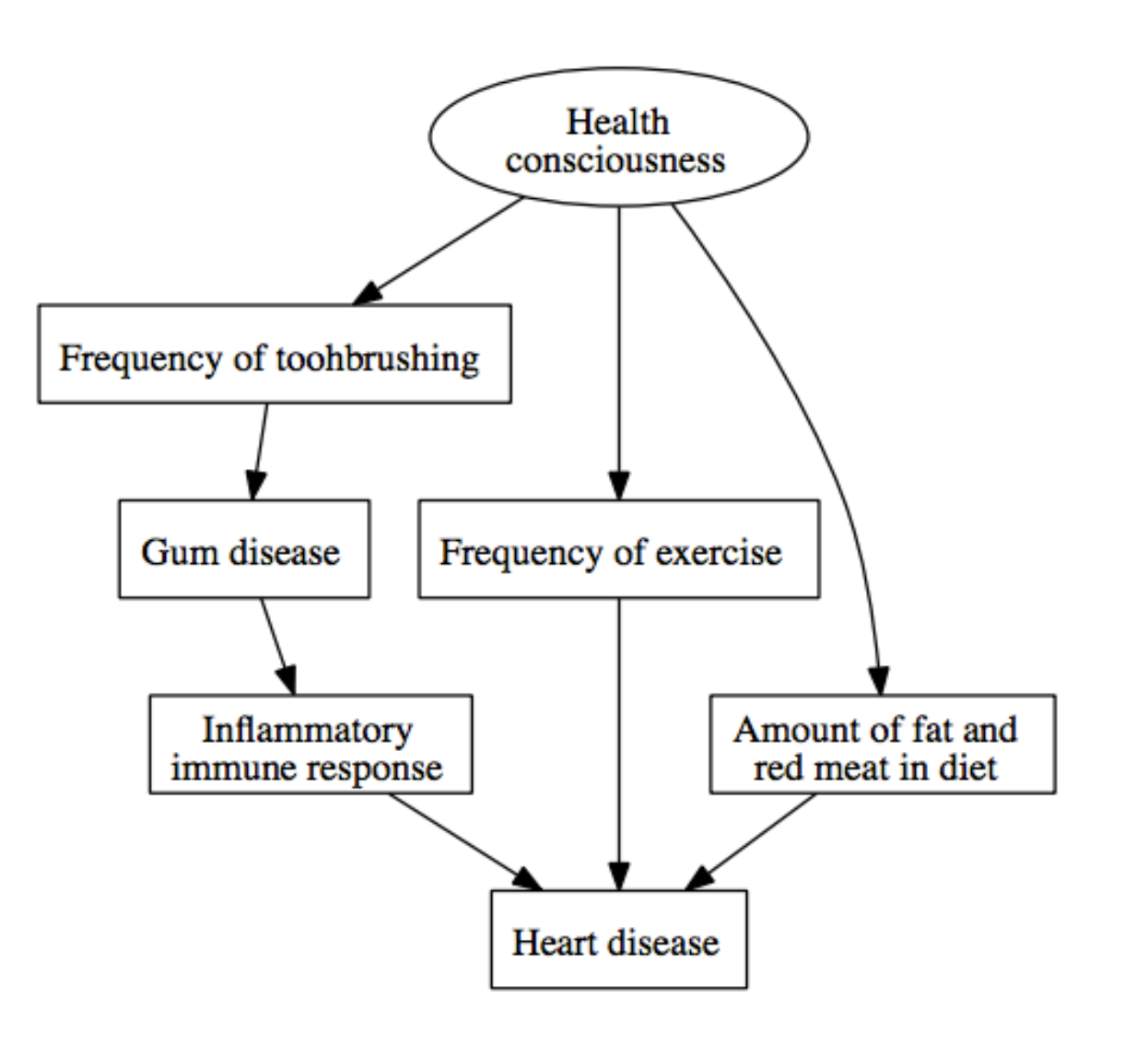
X_i AND Y_i ARE INDEPENDENT, BUT NOT IF WE CONDITION ON Z_i.

Example: Your car won't start either because you ran out of gas or the battery is dead. These are independent events, but if you condition on the car not starting, they are anti-correlated.

CONFOUNDING

- (smoking, cancer) → genetics
- (running the ball, winning games) → having a lead
- (facebook fan, makes purchase) → brand loyalty
- (shown ad, makes purchase) → intent to buy
- (toothbrushing, heart disease) → health consciousness
- (SAT scores, success in college) → genetic intelligence

BIGGER EXAMPLE



WHY CAUSAL INFERENCE?

1. **Science**: Why did something happen?
2. **Decisions**: What will happen if I change something?

SCIENCE

1. Associations are always more interesting when they're causal.
2. Understanding a phenomenon is different than predicting it.

DECISIONS

- quit smoking?
- run the ball more?
- try to recruit Facebook fans for my application?
- purchase costly advertisements?
- brush your teeth? :)
- take an SAT prep class?

A SOCIAL SCIENCE PROBLEM

Causality is easy in natural sciences.

Molecules, cells, animals, plants are exchangeable!

There is always a great deal we don't observe about people.

LINEAR REGRESSION

$$Y_i = \beta X_i + \epsilon_i$$

Often assumes:

$$X_i \rightarrow Y_i \leftarrow \epsilon_i$$

ADD ALL THE CONTROLS YOU WANT

You can always argue that there exists some component of ϵ_i that affects X_i and driving the outcome.

RANDOM ASSIGNMENT

Key idea: enforce that X_i is exogenous by assigning it randomly

Prevents any confounding from being possible.

The gold standard in clinical trials and policy experiments.

MORE REALISTIC:
The Quasi-Experiment

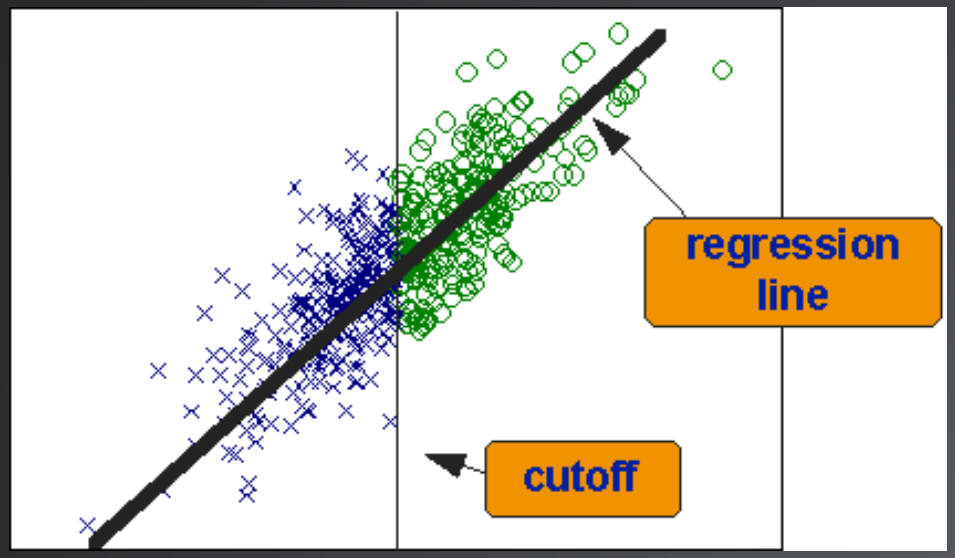
Matching and Opt-In Bias

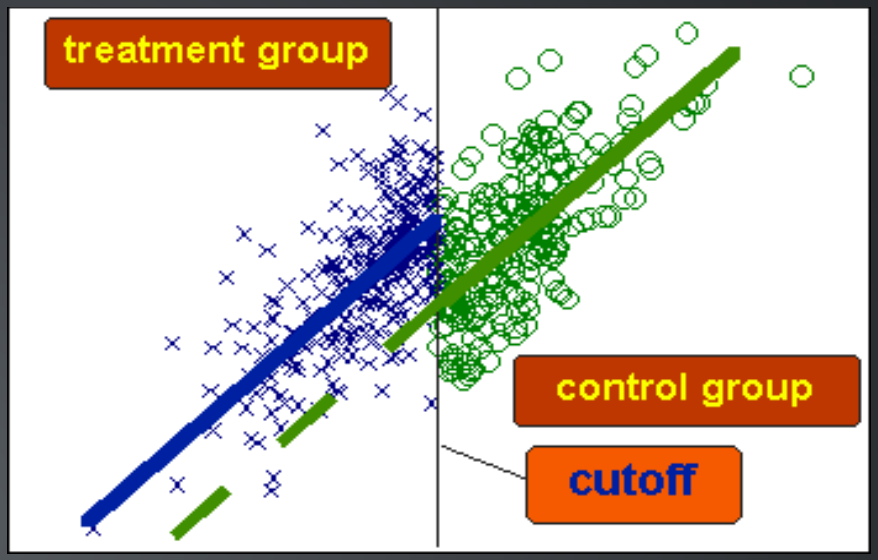
Key idea: make comparisons between observations that are as similar as possible on as many observable dimensions as you can.

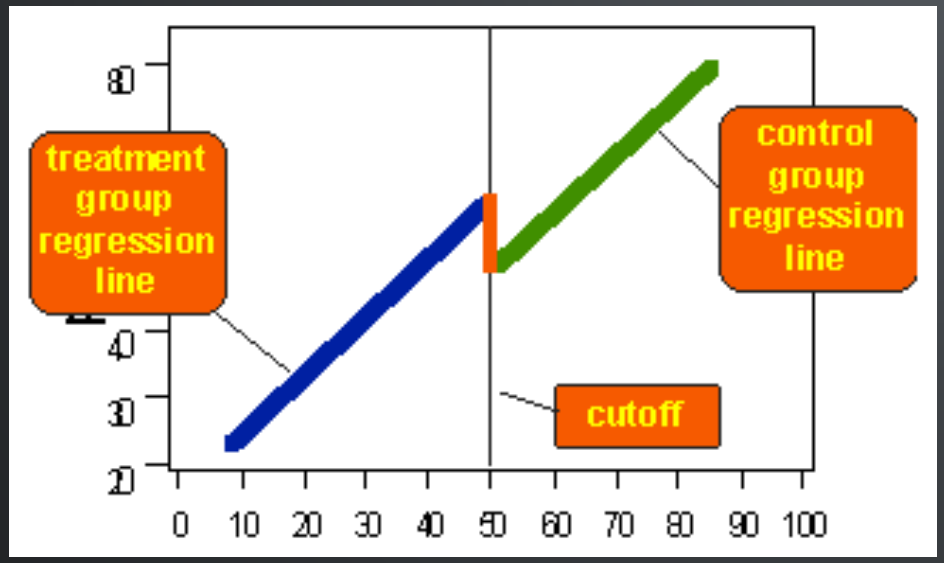
MATCHING

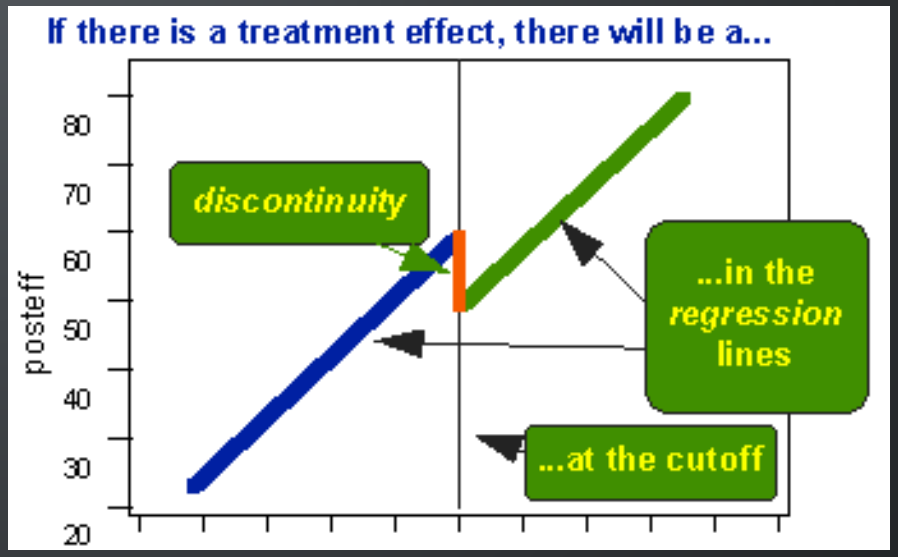
- For every user with an adopter friend, find another user in the population.
- Make sure they are as similar as possible on every characteristic you can measure.
- Discard users for whom you cannot find a suitable match.
- **KEY:**Find characteristics which are good proxies for the latent confounding variables.

REGRESSION DISCONTINUITY



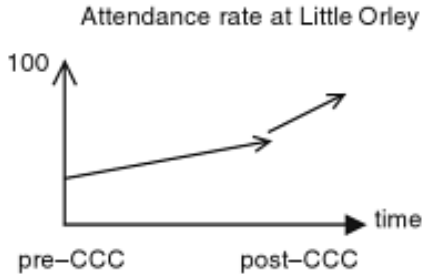




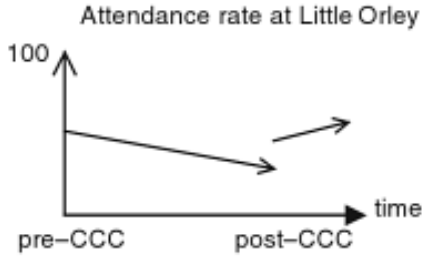


TIME SERIES EVALUATION

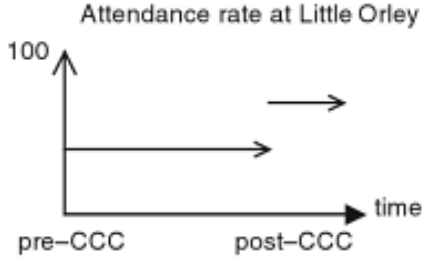
(a) Impact on Slope: Upward Trend; Gradual Impact



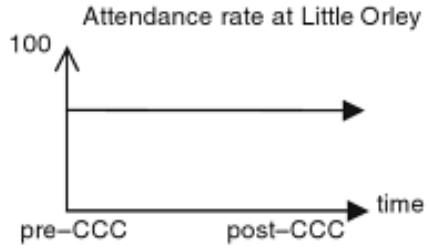
(b) Impact on Slope and Intercept: Downward Trend, Immediate Impact, Reversal of Trend



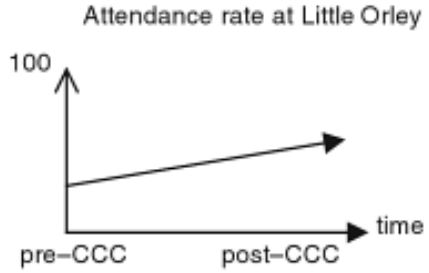
(c) Impact on Intercept: No Trend, Immediate Impact



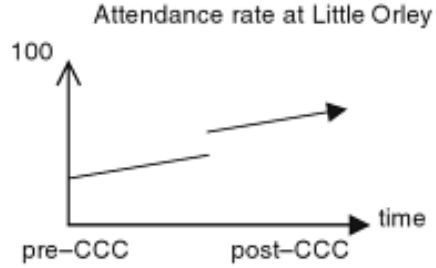
(d) No Impact: No Trend, No Impact



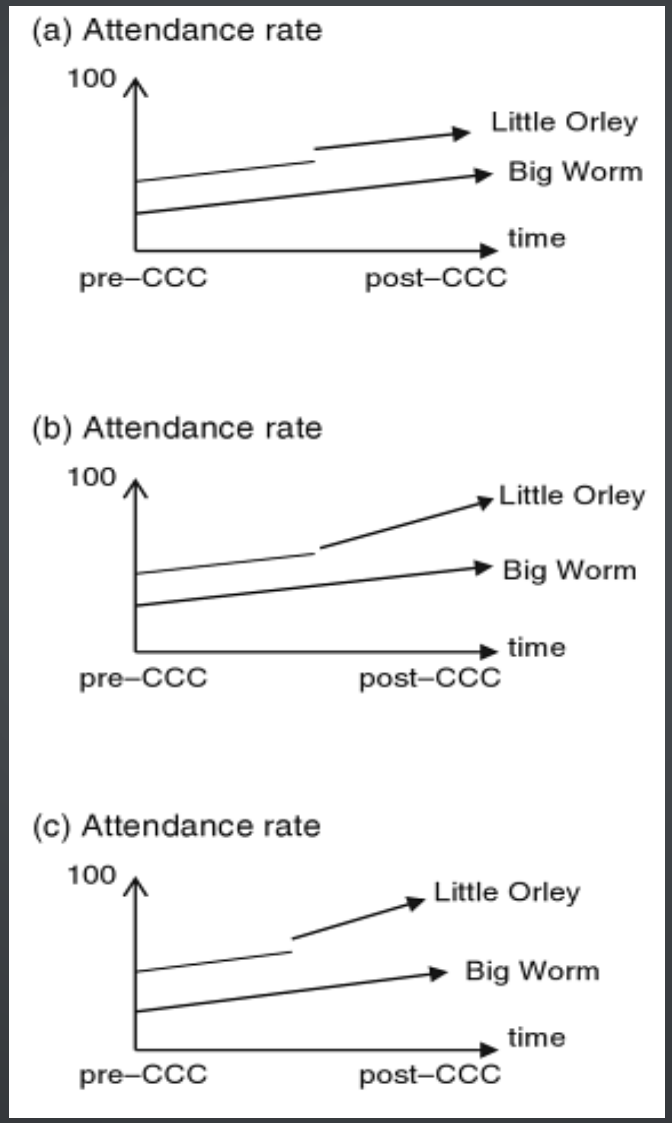
(e) No Impact: Upward Trend, No Impact



(f) Impact on Intercept: Upward Trend, Immediate Impact



TIME SERIES WITH COMPARISON GROUP



QUESTIONS??